Data 583 Life Expectancy (WHO)

Justin Chan, Kenny Tong, Viji Rajagopalan

7 Mar, 2023

EDA

Original Dataset Summary & Initial Data Screening

Purpose : Let's take a snapshot of the original dataset and have a rough idea of its record

```
le <- read.csv("dataset/LifeExpectancy.csv")
summary(le)
```

```
##    Country               Year         Status          Life.expectancy
##  Length:2938        Min.   :2000   Length:2938        Min.   :36.30
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.10
##  Mode  :character   Median :2008   Mode  :character   Median :72.10
##                     Mean   :2008                      Mean   :69.22
##                     3rd Qu.:2012                      3rd Qu.:75.70
##                     Max.   :2015                      Max.   :89.00
##                                                       NA's   :10
##  Adult.Mortality infant.deaths       Alcohol        percentage.expenditure
##  Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000
##  1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685
##  Median :144.0   Median :   3.0   Median : 3.7550   Median :   64.913
##  Mean   :164.8   Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251
##  3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534
##  Max.   :723.0   Max.   :1800.0   Max.   :17.8700   Max.   :19479.912
##  NA's   :10                       NA's   :194
##   Hepatitis.B       Measles               BMI        under.five.deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
##  1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00
##  Median :92.00   Median :    17.0   Median :43.50   Median :   4.00
##  Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mean   :  42.04
##  3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20   3rd Qu.:  28.00
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
##  NA's   :553                        NA's   :34
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.755    Median :93.00   Median : 0.100
##  Mean   :82.55   Mean   : 5.938    Mean   :82.32   Mean   : 1.742
##  3rd Qu.:97.00   3rd Qu.: 7.492    3rd Qu.:97.00   3rd Qu.: 0.800
##  Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600
```

```
##  NA's   :19      NA's   :226       NA's   :19
##       GDP              Population        thinness..1.19.years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.10
##  1st Qu.:   463.94   1st Qu.:1.958e+05   1st Qu.: 1.60
##  Median :  1766.95   Median :1.387e+06   Median : 3.30
##  Mean   :  7483.16   Mean   :1.275e+07   Mean   : 4.84
##  3rd Qu.:  5910.81   3rd Qu.:7.420e+06   3rd Qu.: 7.20
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.70
##  NA's   :448         NA's   :652         NA's   :34
##  thinness.5.9.years Income.composition.of.resources   Schooling
##  Min.   : 0.10      Min.   :0.0000                  Min.   : 0.00
##  1st Qu.: 1.50      1st Qu.:0.4930                  1st Qu.:10.10
##  Median : 3.30      Median :0.6770                  Median :12.30
##  Mean   : 4.87      Mean   :0.6276                  Mean   :11.99
##  3rd Qu.: 7.20      3rd Qu.:0.7790                  3rd Qu.:14.30
##  Max.   :28.60      Max.   :0.9480                  Max.   :20.70
##  NA's   :34         NA's   :167                     NA's   :163
```

Let's look at the dataset dimension first

```
dim(le)
```

```
## [1] 2938   22
```

Then, have a quick overall screening of the dataset

```
#NOTE: might consider to remove this since str(le) provided us same information but in a more presentab
#head(le,5)
```

Here is another view :

```
str(le)
```

```
## 'data.frame':    2938 obs. of  22 variables:
##  $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
```

```
##  $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

From the above broad view, the following Conclusion/Key Findings are reached :

- The records range is from Year 2000 to 2015
- Columns with NA : Life Expectancy, Adult Mortality, Alcohol, Hep B, BMI, Polio, Total exp, Dip, GDP, Population, thinness..1.19, thinness.5.9, Income.composition.of.resources, Schooling
- 'Status' Column is of the "character" data type, with values "Developing" and "Developed". We will introduce a new column 'Status.val' to be the factor value of 'Status' for better analysis..
- 'Percentage Expenditure' has a mean value of 738.2512955 and max. value of 1.9479912×10^4. Spending on health is more than the GDP per capita? Look into the column definition : Expenditure on health as a percentage of Gross Domestic Product per capita(%). The data of such magnitude simply does not quite make sense. Cross check with other references (e.g. the World Bank https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS). OK, let's conclude that we have hesitation about the reliability/interpretation of the value of this column, and probably would drop and skip this column for the rest of this analysis.
- 'Population' and 'GDP' have a relatively large scale, compared with all other columns. So, we may need to scale these two columns.

Now, let's do some data wrangling based on the above conclusions :

```
# Create a new column Status.val to represent the Status column with number
le$Status.val <- ifelse(le$Status == "Developed",1,0)

# Create a new column as the scaled version of the GDP & Population,
le$GDP_scaled = scale(le$GDP)
le$Population_scaled = scale(le$Population)

# Remove the unreliable column
le <- subset(le,select=-c(percentage.expenditure))
```

Null Value Analysis and Handling

```
library(magrittr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

```
le %>% group_by(Country) %>% summarise(COUNT = n())
```

```
## # A tibble: 193 x 2
##    Country             COUNT
##    <chr>               <int>
##  1 Afghanistan            16
##  2 Albania                16
##  3 Algeria                16
##  4 Angola                 16
##  5 Antigua and Barbuda    16
##  6 Argentina              16
##  7 Armenia                16
##  8 Australia              16
##  9 Austria                16
## 10 Azerbaijan             16
## # ... with 183 more rows
```

Purpose : Investigate the and determine how to handle the null value in the data set

Missing values could have a large affect to the overall quality of the static models and machine learning models and need to be clean before using it in our training model.

Lets investigate how many missing values within our features:

```
library(magrittr)
library(dplyr)
library(tidyr)

missing.values <- le %>%
    gather(key = "key", value = "val") %>%
    mutate(is.missing = is.na(val)) %>%
    group_by(key, is.missing) %>%
    summarise(num.missing = n()) %>%
    filter(is.missing==T) %>%
    select(-is.missing) %>%
    arrange(desc(num.missing))

missing.values
```

```
## # A tibble: 16 x 2
## # Groups:   key [16]
##    key                           num.missing
##    <chr>                               <int>
```

```
##  1 Population                          652
##  2 Population_scaled                   652
##  3 Hepatitis.B                         553
##  4 GDP                                 448
##  5 GDP_scaled                          448
##  6 Total.expenditure                   226
##  7 Alcohol                             194
##  8 Income.composition.of.resources     167
##  9 Schooling                           163
## 10 BMI                                  34
## 11 thinness..1.19.years                  34
## 12 thinness.5.9.years                    34
## 13 Diphtheria                            19
## 14 Polio                                 19
## 15 Adult.Mortality                       10
## 16 Life.expectancy                       10
```

There are total of 2563 missing value within our dataset, we could visualize the missing data to identify patterns or cluster of missing values within our data to determine the cause of the missing data and whether it is random or systematic and to highlight potential biases that may exist in our data set. Visualizing the missing value also allow to understand the extend of the missing data and determine appropriate strategies for imputing missing value, since different imputation methods could be more appropriate depending on the pattern of the missing data.

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
missing.values <- le %>%
  gather(key="key", value="val") %>%
  mutate(isna=is.na(val)) %>%
  group_by(key) %>%
  mutate(total=n()) %>%
  group_by(key,total,isna) %>%
  summarise(num.isna=n()) %>%
  mutate(pct=num.isna/total * 100)
```

```
## Warning: attributes are not identical across measure variables; they will be
## dropped
```

```
## `summarise()` has grouped output by 'key', 'total'. You can override using the
## `.groups` argument.
```

```
levels <- (missing.values%>%filter(isna==T) %>% arrange(desc(pct)))$key
```